Some practical aspects of digital transmission

Because of its many advantages, the digital system is now coming into widespread use, not only for the transmission of inherently digital signals, but for the general transmission of all the signals encountered in telecommunication

John R. Pierce *Bell Telephone Laboratories, Inc.*

There are three worlds of digital transmission—digital transmission as an area of academic research, in which theoretical limitations and complicated schemes of modulation and encoding play the starring role; digital transmission in exotic space and military applications, in which it is, or may seem to be, worthwhile to search through the noise in a huge bandwidth in order to receive a few bits per second; and the digital transmission of everyday bulk communication. This article is concerned with that third area, in which apparatus must be economical and reliable; in which the FCC limits the bandwidth we may use in radio transmission; in which the increase in attenuation with frequency limits the bandwidth it is practical to use in cable transmission; in which error rates should be low enough that error correction is an expedient that need be resorted to only in special situations.

In the past, most transmission by wire and radio was analog transmission. In analog transmission, the signal sent over a cable or from antenna to antenna varies smoothly with variations of the original, or baseband, signal to be transmitted. Today, an increasing amount of transmission is digital transmission, in which discrete rather than continuous signals are transmitted.

Digital transmission is ideally suited to sending signals that are inherently discrete, such as teletypewriter signals or any encoding of numbers of letters or words.* However, by means of pulse code modulation, analog signals such as voice signals and television signals can be transmitted over digital transmission systems with any desired degree of accuracy. Further, digital transmission has many advantages. Through the advance of the solid-state art it is increasingly easy to implement digital circuitry cheaply and so that it functions with negligible error. In digital transmission involving many repeaters, it is practical to regenerate the signal at repeater points, so that the effect of noise is not cumulative, though the infrequent errors the noise produces do accumulate. Digital transmission is coming into wide use, not only for transmitting inherently digital signals, but for the general transmission of all the signals encountered in telecommunication.

Today, a digital transmission system called T1 is in wide use in the Bell System.[1] This system transmits three-level pulses at a rate of 1.5 megabits per second over cable pairs. The possibilities of higher-rate digital transmission via coaxial cable have been explored experimentally.[2] Waveguide[3] seems particularly well adapted to digital transmission. At frequencies above 10 GHz,[4] digital transmission may allow more microwave communication in a given geographical area than forms of modulation that are more susceptible to interference. Further, the widespread transmission of data signals over analog systems shows that the need for digital transmission is growing rapidly.

In analog transmission we send a smoothly varying signal, one that can have any of a range of values, and receive a signal that also varies smoothly, but differs from the transmitted signal through the influence of noise, interference, or distortion.

In digital transmission, during a certain period of time we select from among a fixed, finite number of possible signals one or more signals and transmit these. At the receiver we try to recognize exactly which signals were sent. Noise or interference and distortion may make the received signal differ from the transmitted signal enough so that we interpret it as resulting from a transmitted signal other than the one that was sent.

In analog transmission, noise, interference, or distortion results in our receiving a signal a little different from the one that was sent. In digital transmission, noise or interference and distortion can occasionally result in our receiving the wrong signal.

It is in principle possible to approach the limiting rate of transmission or channel capacity defined by Shannon in his book, *A Mathematical Theory of Communication.* To approach this transmission rate one would have to use very long, complicated, random-appearing signals. At the transmitting end we would select among such signals. At the receiving end we would have to compare the received signals with a code book containing very, very many long complicated signals. In practice, this plan is entirely impractical, however efficient it may be in theory.

Digital transmission is usually carried out by sending a

* Sometimes the string of digits that is transmitted over a digital transmission system is an encoding of the actual discrete message by means of an error-correcting code. Within limits, this makes it possible to correct at the receiver errors in transmission of individual digits or groups of digits. In order to attain such error correction, additional check digits must be transmitted along with the message digits.

sequence of short, equally spaced signals, which may be pulses of various discrete amplitudes, or waves of constant amplitude that have one of several allowed phases. Each of the successive signals represents a particular number, or a particular digit of a number, that it is desired to transmit. At the receiving end we look at the amplitude or phase of the received signal at appropriate sampling times, and from the amplitude or phase at such sampling times we decide what number or digit was transmitted. In this article we shall consider particularly transmission by means of pulses having one of a given number of discrete amplitudes.

Inherently, such a process is inefficient compared with Shannon's optimal encoding. Figure 1,[5] for instance, shows calculated values of channel capacity and bit rate for multilevel pulse transmission for a coaxial cable assuming a given size cable, a given length of cable, a given transmitter power, and a given constant noise power density at the receiving point.

For the example of Fig. 1, Shannon's information theory tells us that we should use signals with a bandwidth of about 150 MHz. The transmitted power density of such signals should be almost uniform. Thus, the received power density will be highest at low frequencies, where the attenuation of the cable is low. With such signals we should be able to transmit about 1600 million error-free bits per second. The large number of bits per hertz means, of course, that quite accurate distinctions of amplitude must be made at the receiver.

If we transmit and detect independently a sequence of band-limited pulses of such shape as to give no intersymbol interference, the attainable information rate is considerably lower, even if we assume that we are able to carry out all processes of modulation and detection and sampling of the received signal perfectly. In order to get a usable signal at the receiver we must transmit most of the power at high frequencies, for which the attenuation of the cable is large, and this is partly responsible for the lower information rate.

Transmission over cable by pulses is particularly inefficient if we use only two sorts of pulses, say, pulses with amplitudes $+1$ and -1. We are able to send many such pulses per second and still detect them in the presence of noise, interference, and distortion. But the signaling rate is only one bit per pulse. Most of the power must be transmitted at the upper end of a very broad band, where the transmission loss is great. If we adopt an error rate of one in 10^{12} at each repeater (one error in 10^8 for 10 000 repeaters), then, ideally, we can send around 260 megabits per second by means of binary pulses, using a bandwidth of about 130 MHz.

If we could satisfactorily implement a multilevel system, ideally we would be able to send some 700 megabits per second with the same error rate. Figure 1 shows that the required optimum number of levels would be 52. This rate is almost half the maximum possible rate for error-free transmission whereas the rate for the two-level system is only about one sixth the maximum.

There is an inevitable loss of efficiency in encoding digital signals by means of pulses that are sought and detected individually at the receiver. Nevertheless, practical digital transmission seems to be committed to either this mode of operation or to a related form of phase modulation. We can somewhat improve the efficiency by using multilevel pulses rather than binary pulses, and in some cases this is sure to be done.

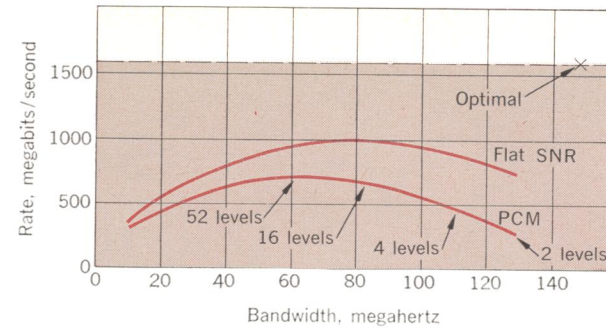


FIGURE 1. Transmission rates in bits/second for a cable whose attenuation in decibels is proportional to $f^{1/2}$, for a prescribed total power and a prescribed flat noise spectrum at the receiving end. "Optimal" is channel capacity for a signal spectrum shaped to give the maximum rate; nothing is transmitted above the frequency indicated. "Flat SNR" is the computed ideal rate for a signal spectrum such that the signal-to-noise ratio is flat at the receiver. The lower curve is for transmission by pulses without intersymbol interference and an error rate of 10^{-12}. (Source: Ref. 5)

We all know that the shorter the pulse the greater the bandwidth. If we are to distinguish the pulses individually at the receiving point, it would appear that we should somehow make the pulses short enough that the tails of one pulse do not overlap the positions at which other pulses are to be received. One solution is to use very short pulses. But this requires an excessive bandwidth.

Nyquist made this clear in a very important paper published in 1928.[6] He showed that there is a particular sort of pulse which can have unit value at time t_0 and which is zero at times t_0 plus or minus $n\tau$, where n is an integer. Such a pulse has a flat amplitude spectrum from frequency 0 to frequency B, and a linear phase. In terms of the bandwidth B, the time τ between successive pulses should be $\tau = 1/2B$. Such a pulse is shown in Fig. 2.

If we send such pulses at intervals τ apart, and if we sample the received signal at appropriate times separated by τ, each sampling will measure only one transmitted pulse, for the tails of all other transmitted pulses are zero at the time corresponding to the center of any chosen pulse. Using such band-limited pulses we can, in principle, completely eliminate intersymbol interference. The lower curve in Fig. 1 was based upon the use of such pulses, which are commonly called sin x/x pulses.

Are sin x/x pulses the only pulses whose tails are zero at all other equally spaced sampling times? The answer is no. There is a whole class of pulses that meet this criterion. And these pulses were also described by Nyquist (see Gibby and Smith[7] for a generalization).

Figure 3 shows a number of sampling times τ apart. It shows two sinusoidal waves. These waves have different frequencies, but each wave has exactly the same value at all sampling points. By merely sampling a wave at intervals τ apart we cannot tell exactly what frequency components are present in the wave.

This ambiguity is illustrated in Fig. 4, which applies to sampling times $1/2B$ apart.

The upper part of Fig. 4 applies if we work with positive frequencies only. If a frequency component f has a

particular sequence of samples, exactly the same sequence of samples can result from a frequency component of frequency $2B - f$, or $2B + f$, or $4B - f$, or $4B + f$, etc. The frequency components that can be the same at all sampling times can be expressed as $2Bn \pm f$.

Sometimes, and especially in connection with complex Fourier transforms of real-time functions, it is desirable to include negative frequencies. In this case a negative frequency component $-f$ always goes along with a positive frequency component $+f$. The set of frequency components that can give the same samples as the pair $+f$, $-f$ are shown in the lower part of Fig. 4. They are $f \pm 2Bn$.

The fact that components of several different frequencies can have the same value at all sampling points makes it possible to construct various pulses that have the desirable property of the sin x/x pulse. That is, one can construct other pulses that are zero at sampling points other than the center of the pulse. Figure 5 shows the spectra of two such pulses. It is assumed that the phase of all frequency components varies linearly with frequency. These spectra meet a criterion given by Nyquist: The sum of the amplitudes of components of frequencies f and $2B - f$ are constant (we assume that no frequencies higher than $2B$ are present).

At times we may wish to shift a pulse upwards in frequency by a carrier frequency f_c and later recover the baseband pulse by multiplying the shifted band of frequencies by the carrier frequency. In such transmission it may be convenient to transmit a little of the lower as well as most of the upper sideband. In demodulating by means of a carrier of frequency f_c, a frequency $f_c - f$ and the frequency $f_c + f$ produce the same output. Figure 6 illustrates another criterion given by Nyquist. A recovered signal will be correct if the sum of transmission at $f_c - f$ and $f_c + f$ is a constant.

If we apply these two criteria to the transmission of pulses shifted in frequency, we find that the condition shown in Fig. 7 will assure perfect transmission. In the vicinity of the carrier frequency, the sum of the transmission at $f_c + f$ and $f_c - f$ is constant. Further, in the vicinity of the band edge $f_c + B$, the sum of the transmission at $f_c + f$ and $f_c + 2B - f$ is constant. It is assumed that the phase is linear. Such a shifted band of frequencies can be demodulated into a baseband pulse that is zero at any instant displaced in time plus or minus $n\pi$ from the center, where τ is equal to $1/2B$.

The sin x/x and other pulses meeting the Nyquist criterion are important in thinking about digital transmission by means of pulses, and they certainly have importance in practical transmission as ideals toward which we can aim. We must, of course, keep in mind that it is in principle physically impossible to generate exact sin x/x pulses or other pulses meeting the Nyquist criterion, for such pulses have infinite tails in both directions and hence would require an infinite delay for their exact synthesis.

Even if we could generate and use independent sin x/x pulses, their employment would lead to certain practical problems. Let us consider, for instance, multichannel digital transmission via radio. The upper curves in Fig. 8 represent the filter characteristics of the combined input and output filters that are necessary to prevent crosstalk in a multichannel radio system. In some ideal sense, these filters could have "square" frequency characteristics, so as to separate the spectrum into contiguous channels of

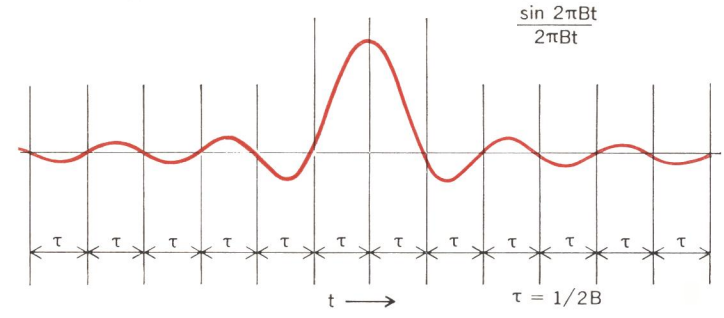


FIGURE 2. A sin x/x pulse.

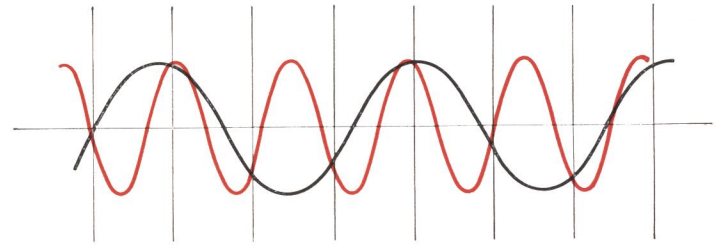


FIGURE 3. Two sine waves of different frequency that have the same values at all sampling times.

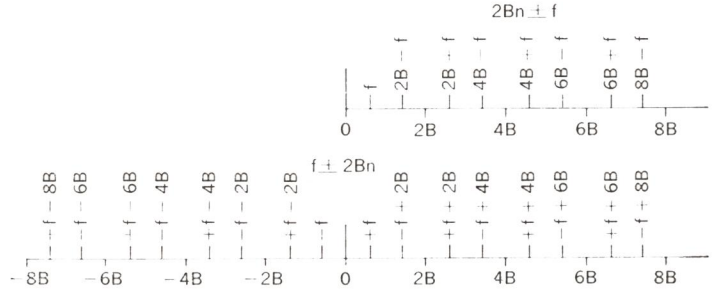
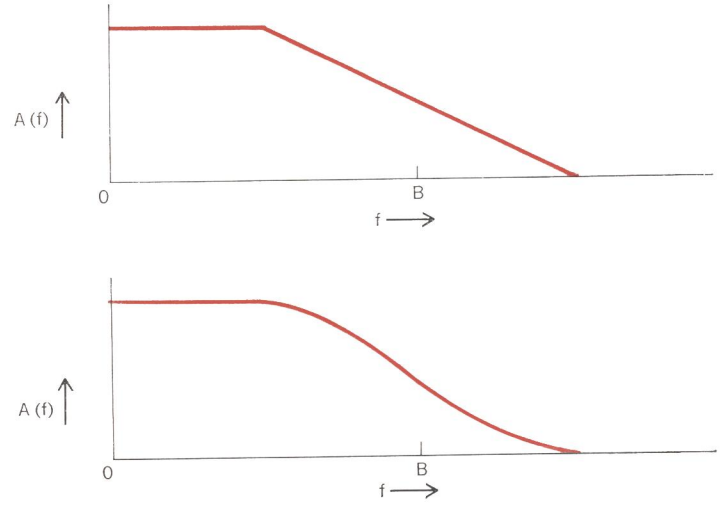


FIGURE 4. Frequencies that can give the same succession of values at sampling points 1/2B apart.

FIGURE 5. Amplitude spectra, which, with linear phase, correspond to pulses meeting the Nyquist criterion for sampling times spaced by $\tau = 1/B$. For these spectra, A(f) + A(2B − f) = constant.



width B_c. However, it is impossible to build such filters and uneconomical to approach them. Thus, whereas transmission can be good near the center of the various rad o channels, the amplitude and the phase of transmission must become increasingly poor toward the edges.

If we try to transmit sin x/x pulses with a flat spectrum of width B over such channels, we can achieve excellent transmission if we make the bandwidth B small compared with the channel spacing B_c. But this means an inefficient use of the spectrum. If we make B larger, the frequency components toward the edge of the band will appear with the wrong amplitude and the wrong phase. Thus, the received signal will not meet the Nyquist criterion. Tails of one pulse will appear at the sampling times of other pulses, and we have intersymbol interference.

Figure 9 shows another difficulty with using a flat spectrum of bandwidth B in transmitting pulses by radio. In order to reconstruct the pulses at the receiver, we need the carrier of frequency f_c. We would like to transmit this carrier. Unfortunately, the flat spectrum extends from f_c to $f_c + B$. If we wish to transmit carrier in any practical way we will have to remove the lower part of the spectrum by filtering, and this will alter the pulse shapes of the demodulated pulses from their ideal sin x/x shape, and so result in intersymbol interference.

Further, we might like to transmit a pilot frequency of frequency $f_c + B$. If we had such a frequency, by taking the differences between this and the carrier frequency we would recover B; by doubling this we have the timing frequency we need in sampling the received pulse train.

We see that the use of a flat spectrum corresponding to sin x/x pulses has disadvantages in radio transmission. It also had disadvantages in coaxial transmission. In principle, wire pairs and coaxial cables provide excellent transmission of direct current. In practice, frequencies near zero cannot be used for several reasons. Among these are earth currents, noise, the transmission of direct current or low-frequency power over the wire pairs or cables, and the practical necessity of using transformers in amplifiers. Thus, we must send over coaxial cables pulses that do not contain any very low frequencies.

In principle, there are several forms of "dc restoration" that might be employed to get the right signal despite the nontransmission of low frequencies.[8,9] For example, in quantized feedback each received or decoded digit is used to produce a transient equal and opposite to the tail or transient following the received pulse corresponding to the digit. This transient is added to the input of the decoder so as to cancel out the tail of the pulse at all subsequent sampling times.

Other remedies have actually been used in coping with the lack of dc transmission. In the T1[1] system an excellent remedy called bipolar transmission[10] was used for digital transmission over twisted pairs.

In bipolar transmission, the digit 0 is transmitted as a pulse of zero amplitude and the digit 1 is alternately transmitted as a pulse of amplitude $+v$ or a pulse of amplitude $-v$. Three levels are used redundantly to transmit a binary signal. We have no choice between the amplitudes $+v$ and $-v$; a pulse of amplitude $+v$ conveys the same information as a pulse of amplitude $-v$. In essence, we use a nonrandom or redundant pulse train to transmit the message digits. Because we use a redundant pulse train, the efficiency of transmission is lowered, but in return the pulse train is given a spectrum which goes to zero at zero frequency.

Thus, the developments in digital transmission have been of the following types.

Because it was impractical to implement the highly efficient sorts of codes that Shannon showed could, in

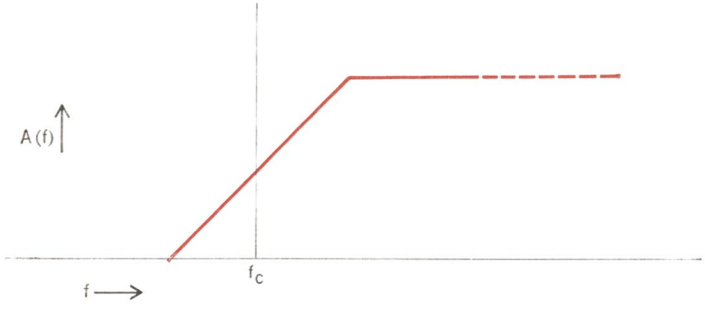


FIGURE 6. Illustration of vestigial sideband criterion $A(f_c - F) + A(f_c + f) = \text{constant}$.

FIGURE 7. Amplitude spectrum of a vestigial sideband pulse that meets the Nyquist criterion.

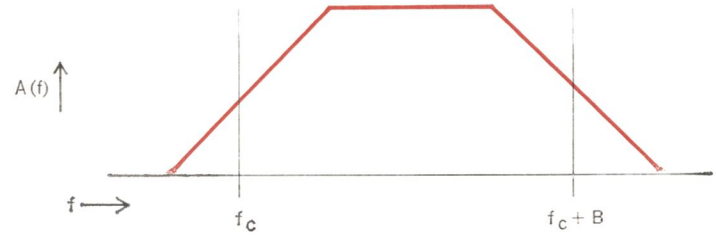
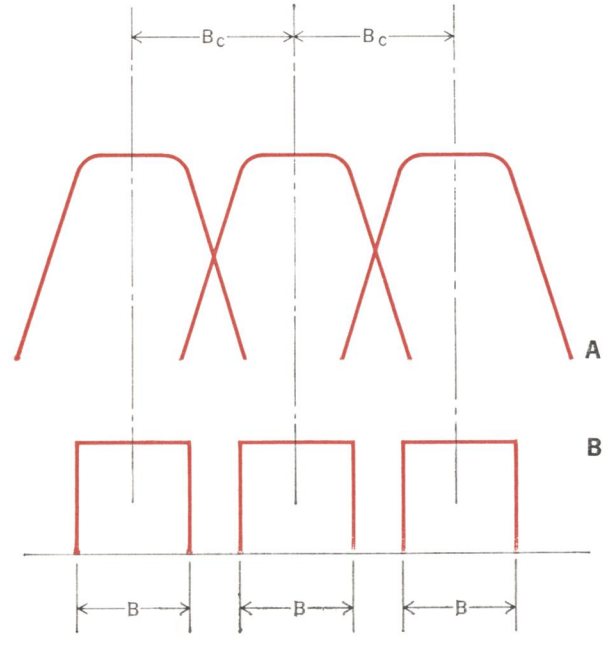


FIGURE 8. Realizable filter characteristics make necessary a greater than "ideal" spacing of channels. A—Attenuation of transmission channels. B—Spectra of SSB sin x/x pulse trains.



principle, realize the full channel capacity of the circuit, people went to the transmission of digital information or other information by means of pulses generated and detected individually. Thus, transmission necessarily falls short of channel capacity of the medium.

Now we go further and send out, not independent pulses whose amplitudes we are completely free to choose, but a redundant pulse train, in order to gain some control over the spectrum of all the pulses transmitted. Thus, we can make the spectrum go to zero where the transmission is poor: near direct current, for instance, in the case of transmission over wire or cable. More generally, by using redundancy in the pulse train we can make the transmitted power go to zero where the transmission is bad in amplitude or phase.

This can be a powerful technique. We used Fig. 9 to illustrate some of the problems of radio transmission. We see that if we can make the spectrum of the frequency-shifted pulses go to zero at the carrier frequency f_c, and at $f_c + B$, we can transmit both the carrier and a pilot frequency $f_c + B$ without damaging the spectrum of the transmitted signal.

Consider also Fig. 8. We see that if the spectrum were not rectangular, but fell to zero at the edges of a band of width B, bad transmission gain and phase near the band edge would not seriously affect the received signal.

How can we control the spectrum of a train of pulses, assuming that each individual pulse has a sin x/x shape, and hence a flat spectrum? The simplest way is called partial response transmission,[11] as shown in the following:

$$0 \rightarrow -1, 0, +1$$

$$1 \rightarrow +1, 0, -1$$

Here we encode a zero of an original binary train as a sequence of three sin x/x pulses, of amplitudes -1, 0, and $+1$, and we encode binary digit 1 as a sequence of three pulses $+1$, 0, and -1. These sequences overlap so that the pulse rate is unchanged. This results in a sequence of pulses that can have three amplitudes -2, 0, or $+2$. In partial response encoding of a binary string of digits we generate and detect pulses one at a time, and we think of the signal that is generated as made up of sin x/x pulses. Actually, a sin x/x pulse cannot appear by itself as a component of the signal. The signal is composed of overlapping pairs of sin x/x pulses.

A sin x/x pulse is shown at the top of Fig. 10. The pair characteristic of partial response encoding is shown below. The spectrum of the pair goes to zero at frequency zero and at frequency B. Its tails are smaller at a given distance from the central region than are the tails of a sin x/x pulse. A pulse stream made up of these paired pulses has three possible amplitudes. A stream of non-redundant three-level pulses could be used to transmit $\log_2 3 = 1.58$ bits/pulse. We use these three-level pulses to transmit $\log_2 2 = 1$ bit per pulse. Through adding redundancy to control the spectrum, we have reduced the efficiency of transmission by a factor 0.63, but we have gained quite a practical advantage in that the spectrum goes to zero at zero frequency and at frequency B.

Partial response encoding has a practical disadvantage. In order to recover the binary stream we must have the correct received pulse amplitudes for the whole past signal. This can be overcome by an expedient of precoding.

The first use of redundancy in order to control the spectrum of the pulse train, and hence to reduce intersymbol interference, was not partial response but the bipolar used in the Tl transmission system, and described earlier. Bipolar encoding differs from partial response encoding in one particular way. Partial response encoding can never result in any frequency component of zero frequency. A positive transmitted pulse is always paired with a negative pulse two pulse positions away. In bipolar encoding, on the average, there are as many positive as negative pulses. However, it is quite conceivable that in bipolar transmission we will transmit all zeros except for a single one. In that case, the only pulse transmitted is a single sin x/x pulse, which has a flat spectrum that does not fall to zero at zero frequency. However, for a random string of input digits, on the average the spectrum of a bipolar string of pulses does fall gradually to zero at zero frequency.

Figure 11 shows, for a process producing a random string of the input digits, the spectrum for partial response encoding and the spectrum for bipolar encoding. The

FIGURE 9. Carrier and pilot frequencies for SSB transmission of trains of sin x/x pulses of bandwidth B.

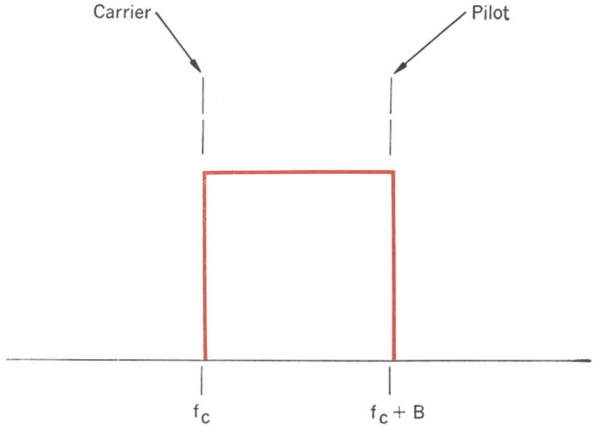
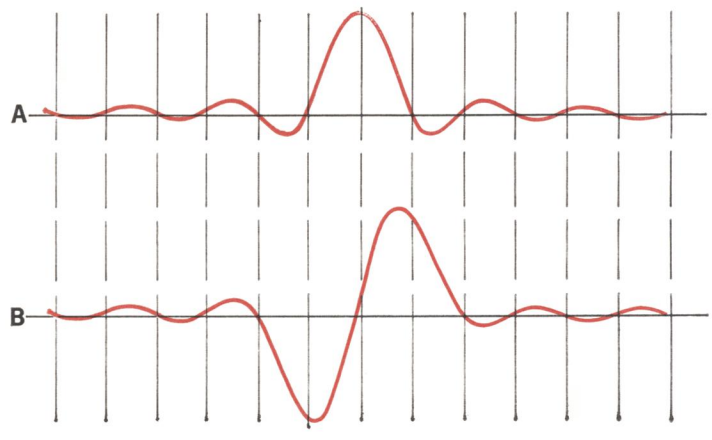


FIGURE 10. Two sin x/x pulses are combined to give the basic waveform used in partial response transmission. A—Independent sin x/x pulses. B—Partial response (pairs of sin x/x pulses). Redundancy , three levels instead of two, $\log_2 2/\log_2 3 = 0.63$.



partial response spectrum goes to zero at zero frequency and at frequency B. It is zero above frequency B because it is assumed that sin x/x pulses of bandwidth B are used. The bipolar spectrum is zero at zero frequency and a maximum at frequency B. It is zero above frequency B because, again, we have assumed sin x/x pulses of bandwidth B.

Actually, the effect of the bipolar encoding is to multiply the spectrum of an individual pulse by a raised cosine function with a period $2B$, which, of course, falls to zero at a frequency $2B$.

Maurice Karnaugh[12] devised an ingenious scheme by means of which any encoding that makes the spectrum go to zero at frequency intervals of $2B$, as bipolar encoding does, can be used to attain a spectrum that goes to zero at intervals B, as partial response encoding does. Figure 12 explains this. At the top is a sequence of digits to be encoded. This stream of $2B$ digits per second is divided into two streams of B digits per second. Each stream is encoded as a bipolar stream (or subjected to some other process) and is turned into pulses of bandwidth B. The two bipolar streams are then interleaved to give the transmitted signal. For a random sequence of input digits, the power spectrum falls to zero at zero frequency and at frequency B.

There are other ways of producing a redundant train of pulses such that the spectrum goes to zero at zero frequency and at a frequency $2B$. Among these are pair-selected ternary.[13] The ingenious trick of Karnaugh can be applied to all such methods to give a spectrum that goes to zero at both zero frequency and at the upper end of the Nyquist band, the frequency B. Hence, we are equipped to make spectra that suit our needs.

There is still, however, the matter of computing the effect of the transmission characteristics on intersymbol interference. If the received and equalized pulse does not meet the Nyquist criterion, the tail of the pulse will be different from zero in many time slots. Intersymbol interference is the summation in a given time slot of tails of pulses assigned to other time slots. How can we calculate intersymbol interference?

If the pulses used have a bandwidth considerably greater than the Nyquist bandwidth, and if the spectrum falls gradually toward zero at the upper end of the band, the pulse tails will be reasonably short and will cause intersymbol interference only in a few nearby time slots. Such pulses are wasteful of bandwidth, however, and are scarcely suitable for efficient radio transmission. These broadband pulses could be used in transmission over cables or wires but there would be a power penalty in transmitting power at unnecessarily high frequencies at which attenuation is high.

If an effort is made to make the pulses broader, and hence more efficient in the sense of conserving frequency, the tails differ from zero at many time slots. If we try to compute intersymbol interference by making a Fourier transform of the spectrum, or by simulation, we will have to examine the sums of many tails in distant time slots. Is there any way of getting the required information simply from the transmission characteristic and the spectrum of the transmitted pulse train?

There is, indeed, a way of computing the *average* intersymbol interference entirely in the frequency domain.[14] This method takes into account the pulse, the characteristics of the transmission circuit, and the modification of the spectrum of the pulse train through the introduction of redundancy by bipolar or some other expedient means. The following exposition was provided by D. Slepian; the form of the result is different from, and perhaps simpler to understand than, that given by Smith. We start by expressing the transmitted signal in terms of a sequence of numbers. Let $\ldots b_{-2}, b_{-1}, b_0, b_1, b_2, \ldots$ be a statistically stationary time series of real numbers that we regard as the stream of digital data to be transmitted.* We use these data to generate a pulse train whose amplitude as a function of time is $s(t)$:

$$s(t) = \sum_{n=-\infty}^{\infty} b_n p(t - n\tau) \qquad (1)$$

Here τ is the pulse spacing interval. The waveform of the individual transmitted pulse $p(t)$ has a Fourier transform $P(f)$, so that

$$p(t) = \int_{-\infty}^{\infty} e^{2\pi jft} P(f)\, df \qquad (2)$$

$$P(f) = \int_{-\infty}^{\infty} e^{-2\pi jft} p(t)\, dt \qquad (3)$$

Thus, the pulse train can also be expressed:

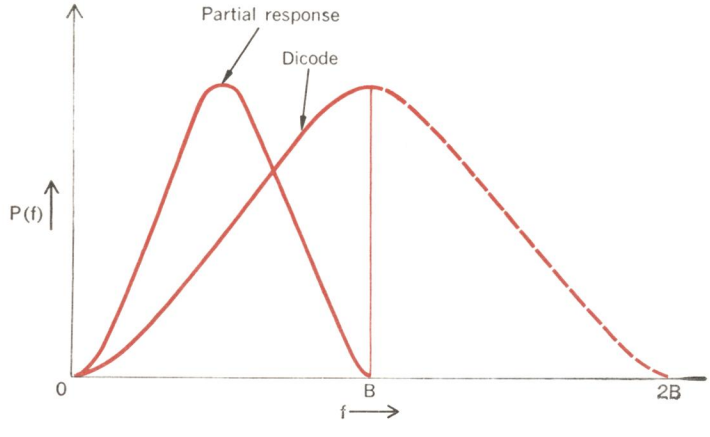
$$s(t) = \sum_{n=-\infty}^{\infty} b_n \int_{-\infty}^{\infty} e^{2\pi jf(t-n\tau)} P(f)\, df \qquad (4)$$

The transmitted signal or pulse train of (1) or (4) is put through a linear transmission system with a transfer function $Y(f)$ that yields a new received signal or pulse train $r(t)$:

$$r(t) = \sum_{n=-\infty}^{\infty} b_n \int_{-\infty}^{\infty} e^{2\pi jf(t-n\tau)} P(f) Y(f)\, df \qquad (5)$$

To recover the data from $r(t)$ we sample this received wave at times $t_0 + n\tau$. Here t_0 is a time displacement that may be adjusted for least intersymbol interference. We multiply the sample amplitudes by a constant factor A so as to make their amplitudes correspond as nearly as

FIGURE 11. Power spectrum densities for partial response and bipolar transmission.



* Although we apply the expressions given to digital or quantized data, the b's need not be quantized and the results apply to samples of a continuous analog signal.

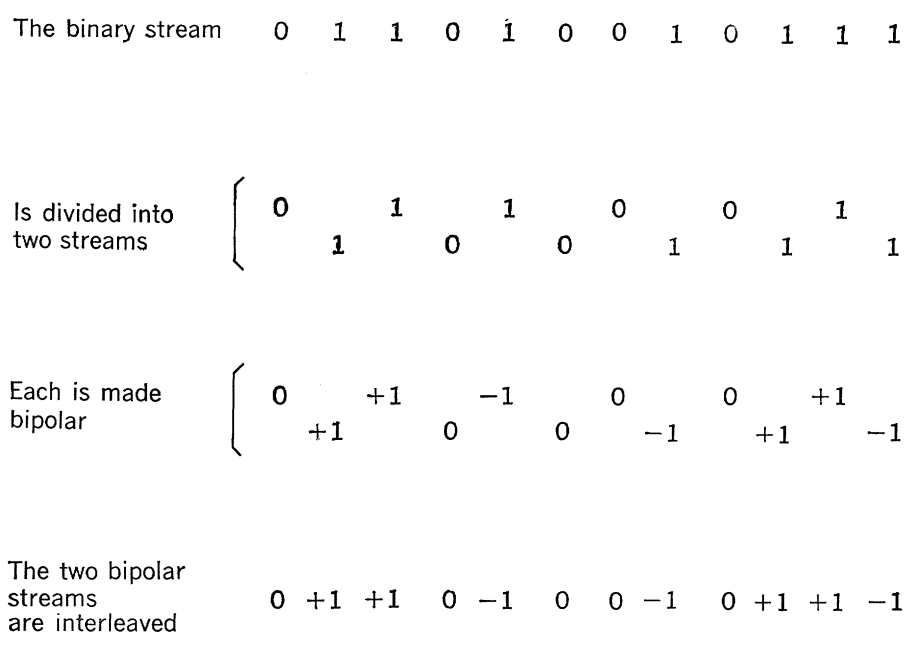

The binary stream
0 1 1 0 1 0 0 1 0 1 1 1
Is divided into two streams
0 1 1 0 0 1 0 1
1 0 0 1 1 1
Each is made bipolar
0 +1 −1 0 0 +1
+1 0 0 −1 +1 −1
The two bipolar streams are interleaved
0 +1 +1 0 −1 0 0 −1 0 +1 +1 −1

FIGURE 12. How to use bipolar transmission and make the spectrum fall gradually to zero at a frequency B.

possible to the coefficients b_n that determine the amplitudes of the transmitted pulses.

It can be shown (from Ref. 14) that the mean square difference ϵ^2 between the samples of the received wave and the coefficients b_n of the pulses of the transmitted wave is given by

$$\epsilon^2 = \tau \int_{-B}^{B} Q(f) \left| 1 - \frac{A}{\tau} T(f) \right|^2 df \qquad (6)$$

$$T(f) = \sum_{n=-\infty}^{\infty} P(f + 2Bn) Y(f + 2Bn) e^{2\pi j t_0 (f + 2Bn)} \qquad (7)$$

In order to attain the least possible intersymbol interference for the transmission system, the constants A and t_0 should be chosen so as to minimize ϵ^2 of (6). This can be done conveniently with the aid of a digital computer.

The quantity $Q(f)$ in (6) is the power spectrum of the data sequence, the sequence of b's. We write for the covariance ρ_l of the data sequence:

$$\rho_l = E b_m b_{m+l} = \rho_{-l} \qquad m, l = 0, \pm 1, \pm 2, \ldots \qquad (8)$$

Here $E b_m b_{m+l}$ means the expected value of $b_m b_{m+l}$. Then

$$Q(f) = \sum_{n=-\infty}^{\infty} \rho_n e^{-2\pi j n \tau f} \qquad (9)$$

The power spectrum $Q(f)$ is periodic with a period $2B = 1/\tau$. For a random sequence of pulses with amplitudes $+1$ or -1, $Q(f)$ is constant and equal to 1. For a bipolar or other sequence $Q(f)$ is not constant, but it is always periodic with a period $2B$. By means of the interleaving trick of Karnaugh we can make $Q(f)$ have a period B rather than $2B$.

The summation of (7) appears because, as we have already seen in connection with Figs. 3 and 4, frequency components $f \pm 2nB$ are indistinguishable at sampling times spaced $\tau = 1/2B$ apart. Hence, we sum over these indistinguishable (at sampling times) components before comparing a received sample with a coefficient b_n.

The quantity $P(f)Y(f)$ represents the complete transmission function of the system excited by the input coefficients b_n, as we see from (5). Quantity $(A/\tau)T(f)$ is a spectral component, adjusted in amplitude by A and in time by t_0. If the Nyquist criterion is met, $T(f)$ can be made equal to unity by properly choosing A and t_0, and hence ϵ^2 can be made to go to zero. We can regard the deviation of $(A/\tau)T(f)$ from unity as a deviation from the Nyquist criterion that for zero intersymbol interference,

$$\sum_{n=-\infty}^{\infty} P(f + 2Bn) Y(f + 2Bn)$$

must have constant amplitude and linear phase.

Finally, $Q(f)$ appears in order to take into account the fact that redundancy in the input sequence of b's can reduce the frequency content of the transmitted and received signals in some frequency ranges and augment it in others.

ϵ^2 is, of course, the mean square error due to intersymbol interference. The distribution of errors is also vitally important. This depends on the nature of the process used in generating the sequence of coefficients b_n that determines the amplitudes of pulses fed into the transmission system. Some work has been done on error bounds due to intersymbol interference,[15] and this important topic deserves further consideration.

Let us now consider the implications and uses of these expressions concerning intersymbol interference.

In designing a transmission system, we will in general expect least intersymbol interference if we use pulses that approximate sin x/x pulses or other pulses whose spectra meet the Nyquist criterion. Such pulses may have spectra considerably broader than the Nyquist bandwidth, B. Exceeding the Nyquist bandwidth may

be acceptable in cable transmission, but it is undesirable in multichannel radio transmission, where spectrum should be conserved.

It is impossible to produce exactly pulses meeting the Nyquist criterion and difficult to approach such pulses. However, with sufficiently complicated equalization it is possible to make intersymbol interference very small. Indeed, virtually all intersymbol interference can be removed by using a self-adjusting equalizer at the receiver[16,17] in which a control circuit adjusts a transversal filter so as to make the response zero at times $\pm n\tau$ from the center of the pulse.

In some cases, a self-adjusting equalizer may be unduly complicated or expensive. A moderate degree of equalization may be practical, but one must use networks of reasonable complexity. In regions where the spectrum is cut off or shaped, the phase will be somewhat nonlinear.

Equation (6) should be a useful guide in arriving at practical networks that will give satisfactory intersymbol interference. The quantity $|1 - (A/\tau)T(f)|^2$ in (6) is a measure of departure of the spectrum from the Nyquist criterion. In some regions this quantity will have an appreciable value. Where this is so, $Q(f)$, the spectrum determined by the redundancy of the pulse train, can be of use. Over some parts of the frequency range we can make $|1 - (A/\tau)T(f)|^2$ small; over other parts we can make $Q(f)$ small. This approach should be a useful resource in reducing average intersymbol interference.

In the exploration that led to this article, the writer's aim was to understand some of the problems of practical efficient digital transmission by radio or cable. He believes that these problems have been reasonably set forth, and that some of the tools useful in exploring and designing such transmission have been discussed. Much more work needs to be done in order to choose and implement a good transmission scheme.

Phase modulation (a form of double-sideband transmission) is particularly attractive for radio and waveguide. Surely, it is important to compare the advantages of phase modulation with those of single-sideband transmission using encodings that make the signal spectrum go to zero in the vicinity of the carrier frequency.

Only a few examples have been given of encodings that make the spectrum of a pulse train go to zero at zero frequency, and at the Nyquist frequency B. It is clear that in all such encodings we must pay a price for shaping the spectrum of the pulse train. Many encoding schemes have been published, and there is surely much unpublished work. Among various coding schemes, which are most efficient? And among the reasonably efficient schemes, which are most practical? A thoroughgoing comparison seems called for.

Further, let us consider the matter of errors. Noise, interference, jitter in the time at which the received wave is sampled, and intersymbol interference can all lead to errors. In radio systems, interfering transmitters may be more important than Gaussian noise. It is not easy to take all sources of error into account at once, even in unreasonably simple cases. Yet, if analysis is to help us, it must be applied to reasonable but somewhat complicated transmission characteristics, whether these be the attenuation of real cables or the transmission characteristics of plausible, practical microwave filters.

In examining the literature, the writer has sometimes encountered papers in which everything (really not everything—usually interference has been left out) is included in complicated expressions, but numerical results are given only for trivial and unrealistic examples, such as RC networks, which afford little insight into actual problems. Might it not be better to explore effects one by one in more realistic cases?

Because we know the probability distribution of Gaussian noise, if we know the signal strength and mean square noise, we can compute the error rate. Presumably, in the case of interfering radio transmitters it should be possible to calculate the error rate in some reasonably straightforward manner. The case of intersymbol interference is more difficult. Whereas a bound has been found and applied to the case of random sin x/x pulses,[15] there is still something to be done.

Having called attention to some of the problems that lie between us and an understanding of what is really possible and practical in digital transmission by radio or cable, the writer will happily leave in more energetic and competent hands the monumental work necessary to resolve these problems.

Essentially full text of a paper presented at a meeting of the Sendai Chapter of the Institute of Electronics and Communication Engineers, Sendai, Japan, April 12, 1968.

REFERENCES

1. Fulz, K. E., and Penick, D. B., "The Tl carrier system," *Bell System Tech. J.*, vol. 44, pp. 1405–1451, Sept. 1965.

2. Dorros, I., Sipress, J. M., and Waldhauer, F. D., "An experimental 224 Mb/s digital repeatered line," *Bell System Tech. J.*, vol. 45, pp. 993–1043, Sept. 1966.

3. Hubbard, W. M., Goell, J. E., Warters, W. D., Standley, R. D., Mandeville, G. D., Lee, T. P., Shaw, R. C., and Clouser, P. L., "A solid-state regenerative repeater for guided millimeter-wave communication systems," *Bell System Tech. J.*, vol. 46, pp. 1977–2018, Nov. 1967.

4. Hogg, D. C., "Millimeter wave communication through the atmosphere," *Science*, vol. 159, pp. 39–46, Jan. 5, 1968.

5. Pierce, J. R., "Information rate of a coaxial cable with various modulation systems," *Bell System Tech. J.*, vol. 45, pp. 1197–1207, Oct. 1966.

6. Nyquist, H., "Certain topics in telegraph transmission theory," *AIEE Trans.*, vol. 47, pp. 617–644, Apr. 1928.

7. Gibby, R. A., and Smith, J. W., "Some extensions of Nyquist's transmission theory," *Bell System Tech. J.*, vol. 44, pp. 1487–1510, Sept. 1965.

8. Bennett, W. R., and Davey, J. R., *Data Transmission*. New York: McGraw-Hill, 1965.

9. Simon, M. K., "Extensions to the analysis of regenerative repeaters with quantized feedback," *Bell System Tech. J.*, vol. 46, pp. 1831–1851, Oct. 1967.

10. Barker, R. H., "Electrical signaling and/or amplifying systems," U.S. Patent 2 700 696, Jan. 25, 1955.

11. Kretzmer, E., "Binary data communication by partial response transmission," Paper G19-65-419, *Conf. Record Cat. 19-C11*, First Ann. IEEE Commun. Conv., Boulder, Colo., June 7–9, 1965, pp. 451–455.

12. Karnaugh, M., "Three-level binary code transmission," U.S. Patent 3 214 749, Oct. 26, 1965.

13. Sipress, J. M., "A new class of ternary pulse transmission plans for digital transmission lines," *IEEE Trans. Communications Technology*, vol. COM-13, pp. 366–372, Sept. 1965.

14. Smith, J. W., "The joint optimization of transmitted signal and receiving filter for data transmission systems," *Bell System Tech. J.*, vol. 44, pp. 2362–2392, Dec. 1965.

15. Saltzberg, B. R., "Intersymbol interference error bounds with application to ideal bandlimited signaling," *IEEE Trans. Information Theory*, to be published.

16. Lucky, R. W., "Automatic equalization for digital communication," *Bell System Tech. J.*, vol. 44, pp. 547–588, Apr. 1965.

17. Lucky, R. W., "Techniques for adaptive equalization of digital communication systems," *Bell System Tech. J.*, vol. 45, pp. 255–286, Feb. 1966.